Hierarchical Pitman-Yor and Dirichlet Process for Language Model

Jen-Tzung Chien and Ying-Lan Chang

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30010, ROC
jtchien@nctu.edu.tw

10.21437/Interspeech.2013-521

Abstract

This paper presents a nonparametric interpretation for modern language model based on the hierarchical Pitman-Yor and Dirichlet (HPYD) process. We propose the HPYD language model (HPYD-LM) which flexibly conducts *backoff smoothing* and *topic clustering* through Bayesian nonparametric learning. The nonparametric priors of backoff n-grams and latent topics are tightly coupled in a compound process. A hybrid probability measure is drawn to build the smoothed topic-based LM. The model structure is automatically determined from training data. A new Chinese restaurant scenario is proposed to implement HPYD-LM via Gibbs sampling. This process reflects the power-law property and extracts the semantic topics from natural language. The superiority of HPYD-LM to the related LMs is demonstrated by the experiments on different corpora in terms of perplexity and word error rate.

Index Terms: language model, backoff model, topic model, Bayesian learning

1. Introduction

Statistical language model (LM) plays an important role in many information systems including machine translation, document classification, writing correction, bio-informatics, and speech recognition. The LM $p(\mathcal{W})$ based on n-gram aims to calculate the probability of a word string $\mathcal{W}$ by multiplying the probabilities of a predicted word w conditional on its preceding $n-1$ words. In general, n-gram model suffers from the inadequacies of training data and long-distance information [7][16]. The modified Kneser-Ney (MKN) LM [6][12] was proposed to tackle the inadequate training data by recursively performing backoff scheme and interpolating with $(n-1)$-grams. The backoff could be also conducted through a structural Bayesian modeling [25]. To compensate insufficient long-distance information, the topic-based language model [8][9][20][23] was constructed by combining large-span latent topic information [1]. An unsupervised LM adaptation was proposed to incorporate topic mixtures based on latent Dirichlet allocation (LDA) [2].

More recently, Bayesian nonparametric (BNP) learning [3] has been extensively studied in machine learning community. BNP methods flexibly infer the model complexity from data without assuming parametric prior and posterior distributions. Teh [22] proposed a BNP approach to backoff LM according to a hierarchical Pitman-Yor (PY) process [15]. Hierarchical PY (HPY) process draws the power-law distributions which is a striking property of natural languages [10]. HPY-LM was interpreted as the *Bayesian extension* of MKN-LM [11][22]. In [19], the class-based HPY-LM was established to characterize many-to-many mapping between words and classes for conversational speech. In [24], a doubly HPY-LM was proposed for LM adaptation. A shared LM was adapted to each domain which was represented by an individual HPY-LM. In [14], a nested HPY-LM was combined with dynamic programming for word segmentation. However, these HPY-LMs [11][14][22][24] did not explore topic information. In [17], a PY topic model was constructed but only for document modeling. In [13], HPY process was combined with a topic model for phrase modeling where the parametric LDA model was considered.

This paper presents the BNP learning for LM which allows model growing structurally as more data are observed. We propose a topic-based LM [5] according to the HPY process compound hierarchical Dirichlet process (HDP) [21]. Using this HPYD-LM, the integrated nonparametric priors are constructed to draw topic-dependent backoff n-grams and simultaneously combine them into a mixture model of topical n-grams. HDP and HPY are tightly integrated to draw HPYD-LM with power-law property and coherent topic information.

2. Prior Works

2.1. Topic-based language model

Topic-based LM [9] was proposed to capture long-range word dependencies through discovery of latent topics. The resulting n-gram is expressed by

$$p(w_i|w_{i-n+1}^{i-1}) = \sum_{z_i} p(w_i|w_{i-n+1}^{i-1}, z_i)p(z_i|w_{i-n+1}^{i-1}) \quad (1)$$

where $z_i = k$ denotes the topic label of word w_i from K topics and $p(z_i|w_{i-n+1}^{i-1})$ denotes the topic proportion given history words $h = w_{i-n+1}^{i-1} = \{w_{i-n+1}, \cdots, w_{i-1}\}$. The topic-based unigrams and bigrams are calculated by $p(w_i) = \sum_{zi} p(w_i|z_i)p(z_i)$ and $p(w_i|w_{i-1}) = \sum_{z_i} p(w_i|w_{i-1}, z_i)p(z_i|w_{i-1})$, respectively. The maximum likelihood estimates of topic-based LM are calculated according to the expectation-maximization (EM) algorithm. In [8], a cache Dirichlet class LM (cDC-LM) was estimated by a variational Bayes EM procedure where the lower bound of log marginal likelihood was maximized. The likelihood was marginalized over latent classes or topics which were represented by Dirichlet distributions. Different from class-based LM [4] based on hard-clustering, the topic-based LMs [8][9] perform soft-clustering over all topics. Nevertheless, these methods calculated the *parametric mixture models* where the number of topics K was fixed. BNP learning aims to relax this assumption and conduct the structural learning.

2.2. Bayesian nonparametric learning

BNP learning using HPY [15][22] and HDP [21] have been proposed to infer LM and document model, respectively. HPY process [22] was developed to draw nonparametric n-gram

25 – 29 August 2013, Lyon, France

model. Given a context U consisting of a sequence of up to $n-1$ history words, HPY-LM calculates the probability of current word w which is sampled from a recursive HPY process $H_U = [H_U(w)]_{w \in \Omega_w}$

$$H_\emptyset \sim \text{PY}(\theta_0, d_0, H_0), \quad H_U \sim \text{PY}(\theta_{|U|}, d_{|U|}, H_{\pi(U)}) \quad (2)$$

where $H_\emptyset$ denotes the word probability over current word w given the empty context $\emptyset$. The global base measure H_0 is viewed as a mean vector given by a uniform value $H_0(w) = 1/|\Omega_w|$ for all vocabulary words $w \in \Omega_w$. The prior process $H_\emptyset = [H_\emptyset(w)]_{w\in\Omega_w}$ draws the unigram $p(w)$. Here, θ_0 and d_0 denote the strength parameter and discount parameter, respectively. In case of $d_0 = 0$, $\text{PY}(\theta_0, d_0, H_0)$ is reduced to $\text{DP}(\theta_0, H_0)$ [22]. Moreover, the probability measure H_U over U is drawn from a PY process based on a prior $H_{\pi(U)}$ from backoff context $\pi(U)$. The strength parameter $\theta_{|U|}$ and discount parameter $d_{|U|}$ depend on the length of context $|U|$. Similarly, the backoff measure $H_{\pi(U)}$ is drawn by the same PY process based on its base measure $H_{\pi(\pi(U))}$ from a even smaller backoff context $\pi(\pi(U))$. Given the unigram probabilities $H_\emptyset$, a *recursive backoff* process is implemented to draw probability measures H_U for bigrams, trigrams, etc. In HPY-LM, many unique words are sampled and most of them rarely. Some frequent but rare words are sampled to meet the rich-gets-richer property [22]. Such power-law property is held due to the discount parameters d_0 and $d_{|U|}$.

On the other hand, HDP deals with the representation of grouped data where each group is associated with a mixture model. Data in different groups share a global mixture model. Each document or group d is drawn from a Dirichlet process (DP) G_d, which determines how much a mixture component from a shared mixture model contributes to that document. The base measure of G_d is itself drawn from a global DP G_g which ensures that a single set of mixtures is shared across documents. The strength parameter θ determines the proportion of a mixture in a document d. The document distribution G_d is generated by

$$G_g \sim \text{DP}(\gamma_0, G_0), \qquad G_d \sim \text{DP}(\theta, G_g) \quad (3)$$

where γ_0 and G_0 denote the strength parameter and base measure of G_g, respectively. HDP is developed to represent "a bag of words" from a set of documents through nonparametric prior G_g. The *sequence of words* is not characterized by HDP.

3. HPYD Language Model

This paper presents a hierarchical Pitman-Yor and Dirichlet language model (HPYD-LM) which jointly conducts backoff smoothing and topic modeling through BNP learning. The number of topics K is learnt from data. This model is different from Bayesian class-based LM $p(w|h) = \sum_{c=1}^{C} p(w|h,c)p(c|h)$ [19] where two HPY processes were developed to separately sample mixture probabilities $p(w|h,c)$ and $p(c|h)$ while number of classes C was fixed. In what follows, HDPY process is shown as a single compound process.

3.1. HPYD process

HPYD-LM assumes that n-gram is expressed by a nonparametric topic mixture model. HPYD process is described as follows. Starting from the uniform seed measure H_0, we draw a global topic measure by $G_g \sim \text{DP}(\gamma_0, H_0)$. The topic-dependent unigram $H_{\emptyset z_i}$ with topic assignment z_i is sampled by $H_{\emptyset z_i} \sim \text{PY}(\theta_1, d_1, G_g)$ where G_g is acted as a prior base

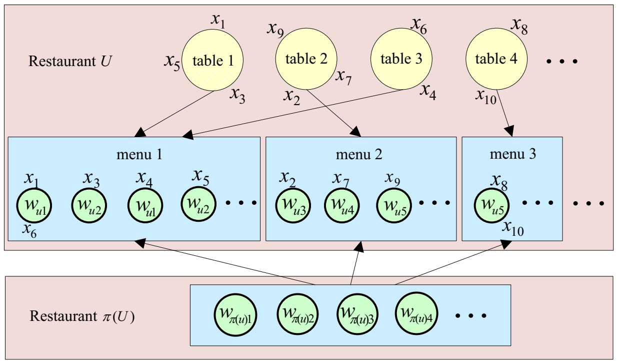

Figure 1: Chinese restaurant scenario for HPYD-LM.

measure. Next, $H_{\emptyset z_i}$ serves as a base measure for a DP to draw unigram probability $G_{w_i} \sim \text{DP}(\gamma_1, H_{\emptyset z_i})$. Using G_{w_i} as a prior measure, we draw topic-dependent bigram by using PY process $H_{w_{i-1}z_i} \sim \text{PY}(\theta_2, d_2, G_{w_i})$. This measure is again acted as a prior basis for a DP to draw bigram $G_{w_i w_{i-1}} \sim \text{DP}(\gamma_2, H_{w_{i-1}z_i})$. Using bigram measure $G_{w_i w_{i-1}}$ as a prior basis, the topic-dependent trigram is drawn by a PY process $H_{w_{i-1}w_{i-2}z_i} \sim \text{PY}(\theta_3, d_3, G_{w_i w_{i-1}})$. Having the prior measure $H_{w_{i-1}w_{i-2}z_i}$, trigram probability is drawn by $G_{w_i w_{i-1} w_{i-2}} \sim \text{DP}(\gamma_3, H_{w_{i-1}w_{i-2}z_i})$. Therefore, HPYD process is recursively realized by sampling topic-dependent n-gram probability $p(w_i|w_{i-n+1}^{i-1}, z_i) \triangleq H_{w_{i-n+1}^{i-1} z_i}$ and then n-gram probability $p(w_i|w_{i-n+1}^{i-1}) \triangleq G_{w_{i-n+1}^{i}}$ by

$$\begin{aligned} H_{w_{i-n+1}^{i-1} z_i} &\sim \text{PY}(\theta_n, d_n, G_{w_{i-n+1}^{i-1}}) \\ G_{w_{i-n+1}^{i}} &\sim \text{DP}(\gamma_n, H_{w_{i-n+1}^{i-1} z_i}) \end{aligned} \quad (4)$$

where $\{\theta_1, \cdots, \theta_n\}$ and $\{d_1, \cdots, d_n\}$ denote the strength and discount parameters of PY process, respectively, and $\{\gamma_0, \cdots, \gamma_n\}$ denote the strength parameters of DP.

3.2. New Chinese restaurant scenario

We implement the nonparametric solution to HPYD-LM in (4) through a new Chinese restaurant process as illustrated in Figure 1. Imagine that there are Chinese restaurants serving customers with infinite tables (yellow) t , infinite menus (blue) k and infinite dishes (green) l. For each restaurant (red) or context U, the first customer or word x_1 enters the restaurant, sits with the first table, and draws a single menu shared for the customers in the same table. He/she orders a dish by this menu. As shown by arrows, tables 1 and 3 draw the same menu 1, table 2 draws menu 2 and table 4 draws menu 3. Let c_{ut} denote the number of customers in table t and n_{ukwl} denote the number of customers ordering dish l which is labelled by a distinct word w from menu k given context U. We have $c_{u\cdot} = \sum_t c_{ut}$, $n_{ukw\cdot} = \sum_l n_{ukwl}$ and $n_{uk\cdot\cdot} = \sum_t c_{u(t=k)}$. Let m_{uk} denote the number of tables that choose menu k and λ_{ukw} denote the number of dishes in menu k which are labelled by distinct word w. Number of occupied tables in restaurant U is expressed by $m_{u\cdot}$. According to this metaphor, each table t is associated with a topic for a distinct menu k and each dish l is associated with an n-gram for a distinct word w. More details are given below.

The ith customer x_i enters a restaurant U and selects either an occupied table with probability $\frac{c_{ut}}{c_{u\cdot}+\gamma_{|U|}}$ or a new table with probability $\frac{\gamma_{|U|}}{c_{u\cdot}+\gamma_{|U|}}$. For a new table, this customer either draws an existing menu k with the probability $\frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma_0}$ or a new menu with probability $\frac{\gamma_0}{m_{\cdot\cdot}+\gamma_0}$. The number of tables drawing menu k in all restaurants $m_{\cdot k}$ is used. Different tables may choose the same menu. After selecting a table with menu k, the customer x_i further selects either an ordered dish l with probability $\frac{\max\{0, n_{ukwl}-d_{|U|}\lambda_{ukw}\}}{n_{uk\cdot\cdot}+\theta_{|U|}}$ or a new dish with probability $\frac{\theta_{|U|}+d_{|U|}\lambda_{u\cdot w}}{n_{u\cdot\cdot\cdot}+\theta_{|U|}}p_{\pi(U)}(w)$. The dishes for $(n-1)$-gram come from back-off restaurant $\pi(U)$ with measure $p_{\pi(U)}(w)$. Number of dishes $\lambda_{u\cdot w}$ is counted over different menus. Combining with new dish provides the approach to model smoothing.

3.3. Gibbs sampling for inference of HPYD-LM

HPYD-LM is inferred according to Gibbs sampling based on this new Chinese restaurant franchise. First, the general topic measure is implemented as a mixture model for K menus (or topics), i.e. $G_g \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma_0}\delta_{\phi_k} + \frac{\gamma_0}{m_{\cdot\cdot}+\gamma_0}H_0$ where $\delta_{\phi_k} \sim H_0$ denotes the atom of topic mixture model for menu k. Next, we draw topic-dependent unigram $H_{\emptyset(z=k)}$ for a word w by considering G_g as a base measure according to the PY process. With the prior measure $H_{\emptyset(z=k)}$, we draw unigram G_{w_i} by a DP. In this fashion, HPYD n-gram with context $U = w_{i-n+1}^{i}$ is recursively sampled by

$$H_{w_{i-n+1}^{i-1}(z_i=k)} \sim \frac{n_{ukw_i\cdot} - d_n\lambda_{ukw_i}}{n_{uk\cdot\cdot} + \theta_n} + \frac{\theta_n + d_n\lambda_{u\cdot w_i}}{n_{u\cdot\cdot\cdot} + \theta_n} G_{w_{i-n+1}^{i-1}}$$
$$G_{w_{i-n+1}^{i}} \sim \sum_{t=1}^{m_{u\cdot}} \frac{c_{ut}}{c_{u\cdot} + \gamma_n} H_{w_{i-n+1}^{i-1}(z_i=t)} + \frac{\gamma_n}{c_{u\cdot} + \gamma_n} H_{w_{i-n+1}^{i-1} z_i} \quad (5)$$

which is realized from (4). In (5), the topic-dependent n-gram $H_{w_{i-n+1}^{i-1}(z_i=k)}$ is first drawn by a PY process with a prior $G_{w_{i-n+1}^{i-1}}$ from backoff context $\pi(U) = w_{i-n+1}^{i-1}$, and subsequently treated as a prior to draw n-gram $G_{w_{i-n+1}^{i}}$ through a DP. Using this HPYD-LM, backoff weights depend on the number of dishes λ_{ukw_i} labelled by word w_i in a menu k. The topic-dependent n-grams are determined through drawing the dishes from different contexts. Topic proportion is decided by the number of customers $c_{u(t=k)}$ sitting in the tables which order the same menu k. Latent topics are autonomously produced by choosing new menus. Therefore, considering the topic-based LM in (1) and the sitting and ordering arrangements of tables, menus and dishes, we infer the nonparametric HPYD-LM $p(w_i|w_{i-n+1}^{i-1})$ which is proportional to

$$\sum_{k=1}^{K} \frac{\sum_t c_{u(t=k)}}{c_{u\cdot} + \gamma_n} \left[\frac{n_{ukw_i\cdot}^{-i} - d_n\lambda_{ukw_i}}{n_{uk\cdot\cdot}^{-i} + \theta_n} + \frac{\theta_n + d_n\lambda_{u\cdot w_i}}{n_{u\cdot\cdot\cdot}^{-i} + \theta_n} \right.$$
$$\left. \times p(w_i|w_{i-n+1}^{i-1}, z_i = k) \right] + \frac{\gamma_n}{c_{u\cdot} + \gamma_n} p(w_i|w_{i-n+1}^{i-1}, z_i = \text{new}). \quad (6)$$

Notably, (6) is viewed as a mixture model consisting of K existing topic-dependent n-grams shown in brackets and a possibly-generated new mixture given in the second term.

We apply the Gibbs sampling algorithm to sample tables, menus and dishes from training data according to the conditional posterior distributions $p(t_i|\mathbf{t}_{-i}, \mathbf{z}, \mathbf{w}, U)$, $p(z_i = k|\mathbf{z}_{-i}, \mathbf{t}, \mathbf{w}, U)$ and $p(l_i|z_i = k, \mathbf{z}_{-i}, \mathbf{w}_{-i}, U)$, respectively, where $\mathbf{w} = \{w_i, \mathbf{w}_{-i}\}$, $\mathbf{t} = \{t_i, \mathbf{t}_{-i}\}$, $\mathbf{z} = \{z_i, \mathbf{z}_{-i}\}$, and "-" denotes the self-exception. The sitting arrangement is determined by sampling table t according to either $p(t_i|\mathbf{t}_{-i}, \mathbf{z}, \mathbf{w}, U)$ which is proportional to $c_{ut}^{-i} \cdot p(w_i|\mathbf{t}_{-i}, \mathbf{z}, \mathbf{w}_{-i}, U)$ if table t is occupied or $\gamma_n \cdot p(w_i|t_i = \text{new}, \mathbf{t}_{-i}, \mathbf{z}, \mathbf{w}_{-i}, U)$ if table t is new. After sitting in a new table, we sample a distinct menu or topic for this table given context U by $p(z_i = k|\mathbf{z}_{-i}, \mathbf{t}, \mathbf{w}, U)$ which is proportional to either $m_{\cdot k} \cdot p(w_i|z_i = k, \mathbf{z}_{-i}, \mathbf{t}, \mathbf{w}_{-i}, U)$ if menu k is ordered or $\gamma_0 \cdot p(w_i|z_i = \text{new}, \mathbf{z}_{-i}, \mathbf{t}, \mathbf{w}_{-i}, U)$ if menu k is new. Next, we draw a dish in menu k by $p(l_i|z_i = k, \mathbf{z}_{-i}, \mathbf{w}_{-i}, U)$ which is proportional to either $\max\{n_{nkw_i\cdot}^{-i} - d_{|U|}\lambda_{ukw_i}, 0\}$ if w_i is ordered or $(\theta_{|U|} + d_{|U|}\lambda_{u\cdot\cdot})G_{\pi(U)}(w)$ if dish $l_i = w$ is new. The counts c_{ut}^{-i} and $n_{nkw_i}^{-i}$ are measured over all words except w_i.

4. Experiments

4.1. Experimental setup

We evaluate the proposed HPYD-LM by using three datasets with different contents and data sizes. The metrics of perplexity and word error rate (WER) (%) are evaluated. In continuous speech recognition, we adopted the Wall Street Journal (WSJ) 1987-1989 corpus containing 86K documents with 38M words and a vocabulary size of 5000. A total of 330 test sentences were sampled from November 1992 ARPA CSR benchmark data. The SI-84 training set was used to estimate HMM parameters based on 39-dimensional MFCC feature vectors. System configuration was detailed in [8]. Two other datasets were collected for evaluation of perplexity. First, the Associated Press newswire (AP) 1989 dataset consisted of 84,778 documents and 1,768,742 sentences with a vocabulary size of 16003. AP was partitioned into a training set with 36,727,591 words and a test set with 4,022,423 words. Second, NIPS0-12 (http://arbylon.net/resources.html) contained 1740 papers from NIPS conferences. We collected a total of 2,034,215 words with a vocabulary size of 3360. NIPS papers were divided into a training set with 1,830,392 words and a test set with 203,823 words.

For comparative study, we carried out trigram LMs by using LDA-LM [20], cDC-LM [8], MKN-LM [6][12] and HPY-LM [11][22]. MKN-LM was carried out by using [18]. The results of LDA-LM and cDC-LM were obtained by interpolating with MKN-LM. HPY-LM and HPYD-LM were implemented by performing Gibbs sampling with 200 iterations and 100 samples at each iteration. The burn-in samples in the first 20 iterations were abandoned. Representative samples from the stationary distribution were collected. The parameters $d_{|U|} \sim \text{Beta}(2, 5)$ and $\theta_{|U|} \sim \text{Gam}(\alpha, \beta)$ were drawn with randomly selected α and β. The parameter $\gamma_{|U|} = 100$ was fixed for all contexts U and n-grams. Perplexity of test data was examined by using SRI toolkit [18] for comparison. We found that the larger the number of words choosing a specific topic, the higher the average discount based on λ_{ukw} is calculated. This phenomenon meets the power law property.

4.2. Experimental results

First of all, Figure 2 displays the estimated topic proportions (number of customers in the tables) in different topics (menus) at different sampling iterations. WSJ corpus is used. The average log probability and the number of estimated topics are shown. Gibbs sampling converges in these iterations. At each iteration, the topics are adaptively estimated. Only a small number of topics have large topic proportions while many topics

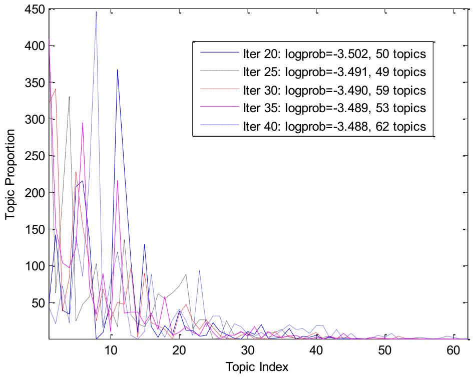

Figure 2: Topic proportions in different sampling iterations.

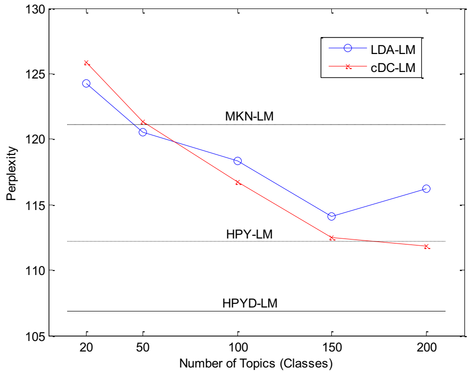

Figure 3: Perplexity versus number of topics using WSJ dataset.

have small topic proportions. Distribution of topic proportions conforms with power-law property. Table 1 shows an example of topic words from five selected topics which are extracted via HPYD-LM using WSJ corpus. It is obvious that topic words within a latent topic are semantically similar while those across topics are significantly different. This topic information is beneficial for estimation of language model. Using HPYD-LM,

Table 1: Topic words using HPYD-LM under different topics.

Trade	Investment	Stock Market	Politics	Economics
million	offer	lynch	purpose	issue
company	previously	plan	public	initial
Tokyo	talk	agency	secretary	information
threat	offer	administration	strike	yen
meeting	interview	bank	Bush	international
transaction	party	dollar	accepted	institute
timing	debt	cents	population	insurance
thrift	investment	cash	aggregate	investment
billion	treasury	Dow	tax	inflation
trade	bank	stock	favorable	income

topics are automatically generated from data. The number of topics is affected by the seed parameter γ_0 as well as the number of n-gram events in training corpus. Larger γ_0 is more likely to draw new menus or lead to a larger topic model. Here, we empirically selected $\gamma_0 = 10$ for three datasets. Using this γ_0, the averaged number of topics in sampling process is 60 for WSJ, 55 for AP and 32 for NIPS. BNP learning works for scalable LMs under different size of training data.

Figure 3 displays the perplexity versus the number of topics or classes by using LDA-LM, cDC-LM, HPY-LM and the proposed HPYD-LM. WSJ corpus is adopted. Using LDA-LM and cDC-LM, the *parametric topic priors* are introduced in Bayesian topic-based LMs where the number of topics is fixed to be K=20, 50, 100, 150 and 200. We can see that perplexity is reduced by increasing number of topics. The lowest perplexity of LDA-LM (114.10) is achieved by using 150 topics. However, BNP learning based on HPY-LM and HPYD-LM reduces the perplexity as 112.20 and 106.84, respectively. Using HPYD-LM, the number of topics (60) is automatically determined. *Nonparametric topic priors* are introduced to build an effective and compact topic-based LM. Furthermore, Table 2 lists the perplexities of MKN-LM, LDA-LM, cDC-LM, HPY-LM and HPYD-LM by using AP and NIPS datasets. HPYD-LM achieves the lowest perplexity among these methods. In case of AP dataset, MKN-LM, LDA-LM, cDC-LM, HPY-LM and HPYD-LM obtain the perplexity of 115.82, 112.35, 109.11, 110.07 and 97.81, respectively. On the other hand, we investigate the performance of continuous speech recognition by using different LMs as shown in Table 3. In this comparison, WERs are reported under comparable model complexity. The results of LDA-LM and cDC-LM were implemented by fixing number of topics or classes to be 60. In this table, HPYD-LMI and HPYD-LM2 imply the HPYD-LM with hyperparameters $\gamma_0 = 1$ and $\gamma_0 = 10$ and lead to 52 and 60 topics in the estimated HPYD-LMs, respectively. In this set of experiments, LDA-LM and cDC-LM are interpolated with MKN-LM and outperform MKN-LM. CDC-LM attains lower WER than LDA-LM and HPY-LM. Among these LMs, the lowest WER 4.82% is achieved by using HPYD-LM with $\gamma_0 = 10$. HPYD-LM could estimate compact LM for speech recognition. The improvement using HPYD-LM comes from the flexibility of backoff smoothing and the contribution of scalable topic information.

Table 2: Perplexity versus different LMs using AP and NIPS datasets.

	MKN-LM	LDA-LM	cDC-LM	HPY-LM	HPYD-LM
AP	115.82	112.35	109.11	110.07	97.81
NIPS	163.18	159.75	155.27	153.02	148.05

Table 3: WER (%) versus different LMs and hyperparameters.

MKN-LM	LDA-LM	cDC-LM	HPY-LM	HPYD-LMI	HPYD-LMII
5.38	5.23	5.10	5.19	4.91	4.82

5. Conclusions

We presented the HPYD-LM based on a new random process which combined HPY for constructing the topic-dependent backoff smoothed LMs and HDP for integrating these LMs into a topic mixture model. Model selection issue was tackled by flexibly extending the number of topics. A Chinese restaurant franchise was proposed to implement the HPYD-LM which satisfied the properties for power-law distribution and topic mixture distribution. The posterior probabilities for drawing tables, menus and dishes were derived. Gibbs sampling was applied to infer HPYD-LM parameters. The experiments on WSJ, AP and NIPS showed that HPYD-LM outperformed the other LMs in terms of perplexity and word error rate.

6. References

[1] Bellegarda, J., "Exploiting latent semantic information in statistical language modeling", *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279-1296, 2000.

[2] Blei, D. M., Ng, A. Y. and Jordan, M. I., "Latent Dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, no. 5, pp. 993-1022, 2003.

[3] Blei, D. M., Griffiths, T. L. and Jordan, M. I., "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies", *Journal of the ACM*, vol. 57, no. 2, article 7, 2010.

[4] Brown, P. F., Pietra, V. J. D., DeSouza, P. V., Lai, J. C. and Mercer, R. L., "Class-based n-gram models of natural language", *Computational Linguistics*, vol. 18, no. 4, pp. 467479, 1992.

[5] Chang, Y.-L. and Chien, J.-T., "Bayesian nonparametric language models", in *Proc. of International Symposium on Chinese Spoken Language Processing*, pp. 188-192, 2012.

[6] Chen, S. F. and Goodman, J., "An empirical study of smoothing techniques for language modeling", *Computer Speech and Language*, vol. 13, pp. 359-394, 1999.

[7] Chien, J.-T., "Association pattern language modeling", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1719-1728, 2006.

[8] Chien, J.-T. and Chueh, C.-H., "Dirichlet class language models for speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 482-495, 2011.

[9] Gildea, D. and Hofmann, T., "Topic-based language model using EM", in *Proc. of European Conference on Speech Communication and Technology*, pp. 2167-2170, 1999.

[10] Goldwater, S., Griffiths, T. L. and Johnson, M., "Interpolating between types and tokens by estimating power-law generators", in *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, 2006.

[11] Huang, S. and Renals, S., "Hierarchical Bayesian language models for conversational speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 1941-1954, 2010.

[12] Kneser, R. and Ney, H. "Improved backing-off for m-gram language modeling", in *Proc. of International Conference on Acoustic, Speech and Signal Processing*, pp. 181-184, 1995.

[13] Lindsey, R. V., "A phrase-discovering topic model using hierarchical Pitman-Yor processes", in *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 214-222, 2012.

[14] Mochihashi, D., Yamada, T. and Ueda, N., "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling", in *Proc. of Annual Meeting of the ACL*, pp. 100-108, 2009.

[15] Pitman, J. and Yor, M., "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator", *Annals of Probability*, vol. 25, pp. 855-900, 1997.

[16] Saon, G. and Chien, J.-T., "Large-vocabulary continuous speech recognition systems - a look at some recent advances", *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 18-33, 2012.

[17] Sato, I. and Nakagawa, H., "Topic models with power-law using Pitman-Yor process", in *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.673-682, 2010.

[18] Stolcke, A., "SRILM-an extensible language modeling toolkit", in *Proc. of International Conference on Spoken Language Processing*, pp. 901-904, 2002.

[19] Su, Y., "Bayesian class-based language models", in *Proc. of International Conference on Acoustic, Speech and Signal Processing*, pp. 5564-5567, 2011.

[20] Tam, Y. C. and Schultz, T., "Dynamic language model adaptation using variational Bayes inference", In *Proc. of Annual Conference of International Speech Communication Association*, pp. 5-8, 2005.

[21] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M., "Hierarchical Dirichlet processes", *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566-1581, 2006.

[22] Teh, Y. W., "A hierarchical Bayesian language model based on Pitman-Yor processes", in *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 985-992, 2006.

[23] Wang, X., McCallum, A. and Wei, X., "Topical n-gram: phrase and topic discovery with an application to information retrieval", in *Proc. of International Conference on Data Mining*, pp. 697-702, 2007.

[24] Wood, F. and Teh, Y. W., "A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation", in *Proc. of International Conference on Artificial Intelligence and Statistics*, vol. 5, pp. 607-614, 2009.

[25] Yaman, S., Chien, J.-T. and Lee, C.-H. , "Structural Bayesian language modeling and adaptation", in *Proc. of Annual Conference of International Speech Communication Association*, pp. 2365-2368, 2007.